様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１８　日現在

機関番号：　１７１０２
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2017
課題番号：　２６２８０００３

研究課題名（和文）文字列情報処理の新展開－文字列組み合わせ論と高度データ構造技術の融合－

研究課題名（英文）New developments in string processing based on combinatorics and advanced data structures

研究代表者

稲永　俊介（Inenaga, Shunsuke）

九州大学・システム情報科学研究院・准教授

研究者番号：６０４４８４０４

交付決定額（研究期間全体）：（直接経費）　　11,700,000円

研究成果の概要（和文）：計算機可読なデータの多くは，記号または文字の羅列，すなわち文字列とみなすことができる．大規模データの効率的な処理・活用を実現するため，本研究では，高速かつ省領域に動作する文字列処理アルゴリズムを開発した．そのために，文字列に内在する組み合わせ的性質を解き明かし，高度データ構造技術を多数開発した．特に，圧縮形式で表現された大規模文字列に対して，高度情報処理を高速かつ省領域に行うアルゴリズムとデータ構造を複数開発した．これらの提案手法は，LZ法，SLP，もしくは RLE と呼ばれる圧縮形式に適用可能である．本研究の成果を，国内・国際会議，国際論文誌において発表し，国内外に発信した．

研究成果の概要（英文）：Digital data can bee seen as sequences of symbols or characters, which are called strings. In this research, we developed fast and space-efficient string processing algorithms that can be a basis for dealing with massive digital data. Combinatorics on words and advanced data structure technologies were the two core primitives for achieving this goal. One of our main contributions is compressed string processing, where the task is to conduct various queries and/or operations over compressed data without explicitly expanding it. Our algorithms work on LZ compression, SLPs, and/or RLE compression. The results obtained in this research have been published in well recognized journals and conference proceedings. We also gave oral presentations in both domestic and international conferences.

研究分野：　理論計算機科学

キーワード：　アルゴリズム　データ構造　文字列組み合わせ論　文字列データ処理

１．研究開始当初の背景

文字列とは，記号の連鎖のことである．コンピュータ上で取り扱うデジタルデータの多くは，文字列とみなすことができる．そのため，文字列データを高速かつ省領域で処理する基盤技術の開発は，情報爆発時代における喫緊の課題となっている．

２．研究の目的

本研究では，文字列データを高速かつ省領域に処理するためのアルゴリズムとデータ構造の開発を行う．文字列照合処理は，文字列処理における代表的な問題であり，以下のように定義される：テキスト T とパターン P の２つの文字列が与えられたとき，T における P の出現位置を求めよ．この亜種として，文字の相対的な順序に着目した順序同型文字列照合問題などがあり，これらの問題を高速かつ省領域で解くためのアルゴリズム技術の開発を行う．また，文字列中に内在する様々な規則性を発見する問題についても取り組む．例えば，同じ文字列の連結であるスクエア（abab など）や，回文（abbba など）およびそれらの拡張・亜種などを探す問題である．

３．研究の方法

近年，爆発的に増加し続けるデータの省領域な格納方法，および有効な活用方法の開発に注目が集まっている．本研究では，文字列組み合わせ論の知見を用いてデータに内在する冗長性を削除し，高速かつ省領域に処理を行う高度データ構造の開発を行う．

４．研究成果

本研究では，実に 39 件の論文を，査読付き国際論文誌および査読付き国際学会会議録にて発表した．以下に，いくつかの代表的な研究成果について述べる．

① 圧縮文字列データからの規則性発見：SLP とよばれる文脈自由文法の一種で表現された圧縮文字列から，回文や反復などの規則的な構造を高速かつ圧縮領域で抽出するアルゴリズムを開発した．提案アルゴリズムは，圧縮文字列を陽に展開することなく処理するため，圧縮サイズに依存したメモリ領域での処理が可能となった．

② 木とグラフに対する順序同型照合：テキストが木またはグラフで与えられた場合の順序同型パターン照合問題を取り扱い，木の場合に高速にパターン照合を行うアルゴリズムを開発した．また，非循環グラフ（DAG）の場合には，この問題が NP 完全であることを示した．

③ 整数アルファベットに対する DAWG の構築：DAWG（Directed Acyclic Word Graph，非循環文字列グラフ）とは，文字列の接尾辞をすべて受理する最小のオートマトンであり，パターン照合問題などに有用なデータ構造である．入力文字列長の多項式サイズのアルファベットの場合において，DAWG 線形時間で構築するアルゴリズムを開発した．

④ 連長圧縮に基づく漸増的編集距離計算：文字列 A と B の編集距離とは，A を B に変換するために必要な挿入・削除・置換の最小回数のことである．動的計画表を用いることで，編集距離を効率よく計算できる．本研究では，文字列 A の先頭に新しい文字が追加される漸増的計算を省領域で行う手法を与えた．連長圧縮文字列に対する知見を活用して解を導いた．

⑤ 簡潔 van Emde Boas データ構造：van Embe Boas データ構造（vEB）とは，1977 年に van Emde Boas が提案した動的な整数データ構造である．vEB は様々なクエリ・操作を高速に処理できるが，膨大な領域を要するという欠点があった．本研究では，簡潔データ構造技術を用いて，省領域 vEB の開発に成功した．

これらの研究成果，また，次節で述べる発表論文の成果は，文字列に内在する組み合わせ的性質を解き明かし，高度データ構造技術を用いることによって得られたものである．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 49 件）（すべて査読付き）

① Tomohiro I, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, Faster Lyndon factorization algorithms for SLP and LZ78 compressed text, Theoretical Computer Science, 656(B):215-224, 2016. 10.1016/j.tcs.2016.03.005

② Yoshiaki Matsuoka, Takahiro Aoki, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, Generalized pattern matching and periodicity under substring consistent equivalence relations, Theoretical Computer Science, 656(B):225-233, 2016. 10.1016/j.tcs.2016.02.017

③ Golnaz Badkobeh, Hideo Bannai, Keisuke Goto, Tomohiro I, Costas S. Iliopoulos, Shunsuke Inenaga, Simon J. Puglisi, and Shiho Sugimoto, Closed

Factorization, Discrete Applied Mathematics, 212:23-29, 2016. 10.1016/j.dam.2016.04.009

④ Kazuyuki Narisawa, Hideharu Hiratsuka, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, Efficient Computation of Substring Equivalence Classes with Suffix Arrays, Algorithmica, 2016. 10.1007/s00453-016-0178-z

⑤ Heikki Hyyrö, Kazuyuki Narisawa, and Shunsuke Inenaga, Dynamic Edit Distance Table under a General Weighted Cost Function, Journal of Discrete Algorithms, 34:2-17, 2015. 10.1016/j.jda.2015.05.007

⑥ Yuto Nakashima, Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Constructing LZ78 Tries and Position Heaps in Linear Time for Large Alphabets, Information Processing Letters, 115(9):655-659, 2015. 10.1016/j.ipl.2015.04.002

⑦ Tomohiro I, Takaaki Nishimoto, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, Compressed automata for dictionary matching, Theoretical Computer Science, 578:30-41, 2015. 10.1016/j.tcs.2015.01.019

⑧ Tomohiro I, Wataru Matsubara, Kouji Shimohira, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, Kazuyuki Narisawa, and Ayumi Shinohara, Detecting regularities on grammar-compressed strings, Information and Computation, 240:74-89, 2015. 10.1016/j.ic.2014.09.009

⑨ Tenma Nakamura, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Order preserving pattern matching on trees and DAGs, Proc. SPIRE 2017, LNCS 10508, pp.271-277, 2017. 10.1007/978-3-319-67428-5_23

⑩ Shintaro Narisada, Diptarama, Kazuyuki Narisawa, Shunsuke Inenaga, and Ayumi Shinohara, Computing longest single-arm-gapped palindromes in a string, Proc. SOFSEM 2017, pp. 375-386, 2017. 10.1007/978-3-319-51963-0_29

⑪ Yuta Fujishige, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Computing DAWGs and Minimal Absent Words in Linear Time for Integer Alphabets, Proc. MFCS 2016, 38:1-38:14, 2016. 10.4230/LIPIcs.MFCS.2016.38

⑫ Takuya Mieno, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Shortest Unique Substring Queries on Run-Length Encoded Strings, Proc. MFCS 2016, 69:1-69:11, 2016. 10.4230/LIPIcs.MFCS.2016.69

⑬ Takaaki Nishimoto, Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Fully dynamic data structure for LCE queries in compressed space, Proc. MFCS 2016, 72:1-72:15, 2016. 10.4230/LIPIcs.MFCS.2016.72

⑭ Takuya Takagi, Shunsuke Inenaga, Kunihiko Sadakane, and Hiroki Arimura, Packed Compact Tries: A Fast and Efficient Data Structure for Online String Processing, Proc. IWOCA 2016, LNCS 9843, pp. 213-225, 2016. 10.1007/978-3-319-44543-4_17

⑮ Takuya Takagi, Shunsuke Inenaga, and Hiroki Arimura, Fully-online construction of suffix trees for multiple texts, Proc. CPM 2016, 22:1-22:13, 2016. 10.4230/LIPIcs.CPM.2016.22

⑯ Yoshiaki Matsuoka, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, and Florin Manea, Factorizing a string into squares in linear time, Proc. CPM 2016, 27:1-27:12, 2016. 10.4230/LIPIcs.CPM.2016.27

⑰ Paweł Gawrychowski, Tomohiro I, Shunsuke Inenaga, Dominik Köppl, and Florin Manea, Efficiently Finding All Maximal α-gapped Repeats, Proc. STACS 2016, pp.39:1-39:14, 2016. 10.4230/LIPIcs.STACS.2016.39

⑱ Heikki Hyyrö and Shunsuke Inenaga, Compacting a dynamic edit distance table by RLE compression, Proc. SOFSEM 2016, LNCS 9587, pp. 302-313, 2016. 10.1007/978-3-662-49192-8_25

⑲ Yoshiaki Matsuoka, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, Semi-dynamic compact index for short patterns and succinct van Emde Boas tree, Proc. CPM

2015, LNCS 9133, pp. 355-366, 2015. 10.1007/978-3-319-19929-0_30

⑳ Hideo Bannai, Tomohiro I, Shunsuke Inenaga, Yuto Nakashima, Masayuki Takeda, and Kazuya Tsuruta, A new characterization of maximal repetitions by Lyndon trees, Proc. SODA 2015, pp. 562-571, 2015. 10.1137/1.9781611973730.38

㉑ Yuto Nakashima, Takashi Okabe, Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Inferring strings from Lyndon factorization, In Proc. MFCS 2014, LNCS 8635, pp. 565-576, 2014. 10.1007/978-3-662-44465-8_48

㉒ Tomohiro I, Shiho Sugimoto, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Computing Palindromic Factorizations and Palindromic Covers On-line, Proc. CPM 2014, LNCS 8486, pp. 150-161, 2014. 10.1007/978-3-319-07566-2_16

㉓ Yasuo Tabei, Hiroto Saigo, Yoshihiro Yamanishi, Simon J. Puglisi, Scalable partial least squares regression on grammar-compressed data matrices, Proc. KDD 2016, pp. 1875-1884, 2016. 10.1145/2939672.2939864

㉔ Djamal Belazzougui, Patrick Cording, Simon J. Puglisi, Yasuo Tabei, Access, rank, and select in grammar-compressed strings, Proc. ESA 2015, pp. 142-154, 2015. 10.1007/978-3-662-48350-3_13

他 25 件.

〔学会発表〕（計 37 件）（すべて国際会議）

① Tenma Nakamura, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Order preserving pattern matching on trees and DAGs, 24th International Symposium on String Processing and Information Retrieval (SPIRE 2017), 2017.

② Hideo Bannai, Shunsuke Inenaga, and Dominik Köppl. Computing All Distinct Squares in Linear Time for Integer Alphabets, 28th Annual Symposium on Combinatorial Pattern Matching (CPM 2017), 2017.

③ Yuta Fujishige, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Computing DAWGs and Minimal Absent Words in Linear Time for Integer Alphabets, 41st International Symposium on Mathematical Foundations of Computer Science (MFCS 2016), 2016.

④ Takuya Takagi, Shunsuke Inenaga, Kunihiko Sadakane, and Hiroki Arimura, Packed Compact Tries: A Fast and Efficient Data Structure for Online String Processing, 27th International Workshop on Combinatorial Algorithms (IWOCA 2016), 2016.

⑤ Paweł Gawrychowski, Tomohiro I, Shunsuke Inenaga, Dominik Köppl, and Florin Manea, Efficiently Finding All Maximal α-gapped Repeats, 33rd Symposium on Theoretical Aspects of Computer Science (STACS 2016), 2016.

⑥ Hideo Bannai, Tomohiro I, Shunsuke Inenaga, Yuto Nakashima, Masayuki Takeda, and Kazuya Tsuruta, A new characterization of maximal repetitions by Lyndon trees, ACM-SIAM Symposium on Discrete Algorithms 2015 (SODA 2015), 2015.

他 31 件

〔図書〕（計 1 件）

① Shunsuke Inenaga, Kunihiko Sadakane, and Tetsuya Sakai, String Processing and Information Retrieval (SPIRE 2016), LNCS 9954, Springer, 2016, 273 pages.

〔その他〕
ホームページ等
https://str.i.kyushu-u.ac.jp/~inenaga/index-j.html

６．研究組織
(1)研究代表者
稲永 俊介（INENAGA, Shunsuke）
九州大学，システム情報科学研究院，
准教授
研究者番号：60448404

(2)研究分担者
坂内 英夫（BANNAI, Hideo）
九州大学，システム情報科学研究院，
准教授
研究者番号：20323644

田部井 靖生（TABAI, Yasuo）

国立研究開発法人理化学研究所，
革新知能統合研究センター（AIP），
ユニットリーダー
研究者番号：20589824